EVS-assisted joint Deblurring, Rolling-Shutter Correction and Video Frame Interpolation through Sensor Inverse Modeling

Rui Jiang*, Fangwen Tu*, Yixuan Long, Aabhaas Vaish, Bowen Zhou, Qinyi Wang
Wei Zhang, Yuntan Fang, Luis Eduardo Garcia Capel, Bo Mu, Tiejun Dai, Andreas Suess

OMNIVISION

*Equal contribution

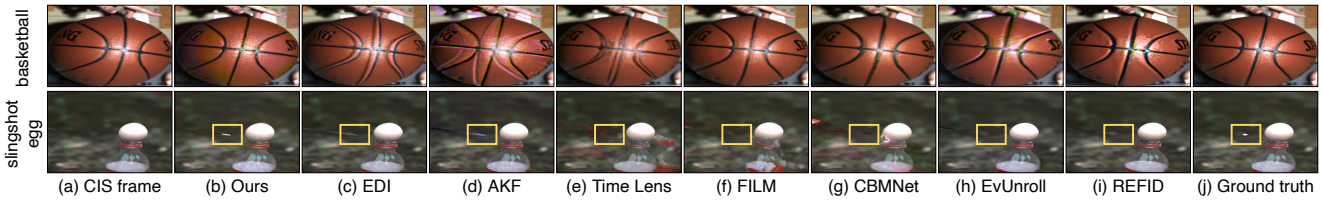

Figure 1. Qualitative results comparing the proposed method with EDI [27], AKF [39], Time Lens [37], FILM [32], CBMNet [23], EvUnroll [44], and REFID [36] on simulation datasets. In the "basketball" scene, our method provides frames that are rolling shutter effect corrected and deblurred. In the "slingshot egg" scene, our method can reconstruct the shape of the pellet with the best quality, as indicated in the yellow box. The images have been magnified for optimal viewing quality.

Abstract

Event-based Vision Sensors (EVS) gain popularity in enhancing CMOS Image Sensor (CIS) video capture. Nonidealities of EVS such as pixel or readout latency can significantly influence the quality of the enhanced images and warrant dedicated consideration in the design of fusion algorithms. A novel approach for jointly computing deblurred, rolling-shutter artifact corrected high-speed videos with frame rates up to $10\,000\,\mathrm{FPS}$ *using inherently blurry rolling shutter CIS frames of* $120\,\mathrm{FPS}$ *to* $150\,\mathrm{FPS}$ *in conjunction with EVS data from a hybrid CIS-EVS sensor is presented. EVS pixel latency, readout latency and the sensor's refractory period are explicitly incorporated into the measurement model. This inverse function problem is solved on a per-pixel manner using an optimization-based framework. The interpolated images are subsequently processed by a novel refinement network. The proposed method is evaluated using simulated and measured datasets, under natural and controlled environments. Extensive experiments show reduced shadowing effect, a* $4\,\mathrm{dB}$ *increment in PSNR, and a* $12\,\%$ *improvement in LPIPS score compared to state-of-the-art methods.*

1. Introduction

Event-Based Vision Sensors (EVS) capture our world from a completely different perspective compared to CMOS Image Sensors (CIS) [9, 14, 24, 25]. In EVS, each pixel independently triggers an event if a relative illuminance change on the pixel reaches a certain triggering threshold (typically in the range of $10\,\%$ to $30\,\%$). Conversely to CIS, EVS operate time-continuously and asynchronously - there is no common exposure period or "frame rate". Thus, EVS pixels can be read out comparatively fast in a low-power manner enabling efficient high-speed data capture. Recent work on enhancement of high-quality images [23, 27, 34, 37, 39] focuses on fusing CIS and EVS information so that the high fidelity of CIS measurement and the fast response of EVS measurement can complement each other.

In practice it is observed [3, 14, 21, 25] that the dynamic characteristics of EVS pixels are highly dependent on the illuminance in the sensor plane. Under low-light conditions such as indoor capture, a slower EVS response during video enhancement can result in ghosting or blurry frames and needs to be considered in algorithm design. Furthermore, the inevitable readout latency and refractory period of EVS pixels can significantly aggravate artifacts. Current methods are not specifically designed to address sensor nonidealities.

This paper presents an inverse model capturing the EVS pixel latency (namely the V_{FE} delay), the readout latency, and the refractory period. The inverse model is solved through nonlinear graph optimization to jointly address deblurring, rolling-shutter (RS) correction and video-frame interpolation while compensating for sensor nonidealities. In addition to the inverse model, a learning-based refinement module is proposed to enhance the image quality by further mitigating the noise and artifacts. The proposed method is compared with existing CIS-based [32] and EVS-assisted [8, 23, 27, 36, 37, 39] video enhancement methods under challenging indoor scenarios with fast motion. As illustrated in Figure 1, the proposed algorithm produces excellent image quality with much fewer artifacts.

The contributions of the paper can be summarized as:

- A formulation of a joint debluring, rolling-shutter correction and video-frame interpolation problem based on CIS and EVS fusion, where EVS sensor nonidealities are explicitly modeled (Section 3),
- An optimization methodology jointly solving this coupled inverse problem (Section 4.1), as well as a post-processing network for artifact removal (Section 4.2),
- A comprehensive comparison and analysis of simulation and measured datasets in natural and controlled environments (Section 5).

2. Related Work

Video Frame Interpolation Video interpolation aims to synthesize intermediate frames between two images to achieve a temporally coherent video sequence. Traditional CIS-based methods depend on optical flow for a smooth transition. In SuperSlomo [18], the bi-directional flows are combined with a linear combination followed by a refinement using U-Net [33]. Such optical flow-based methods, however, suffer from occlusion problems. In DAIN [1], researchers estimate occlusion areas using depth information and use contextual information from neighboring pixels to fill the occluded regions. In RIFE [17], the flow and fusion maps are generated simultaneously without an additional optical flow module for final result synthesis. FILM [32] presents a multi-scale feature extractor within a unified network to achieve an enlarged receptive field aimed at handling large object motion in the video.

Event-assisted Frame Enhancement A key issue lowering the image quality of CIS-based frame interpolation methods is the lack of information between frames. EVS provides such information thanks to its asynchronous low latency capture not relying on a global exposure period. Some early studies [31, 34] show that intensity reconstruction purely based on EVS modality is feasible. However, as EVS captures events based on differential changes of illuminance, this approach struggles to estimate the absolute light level. In [2], the optical flow and intensity frames are estimated simultaneously by solving an optimization problem that takes EVS measurements as data terms. The authors in [22] take a step further by formulating the event-based SLAM problem, which can be decomposed into camera ego-motion estimation and 3-D scene reconstruction. However, the quantization error makes it difficult to reconstruct high-quality frames from EVS alone. As EVS pixels trigger independently, researchers simplify the problem formulation to pixel-wise estimation by fusing CIS and EVS, without seeking explicit representation of the 3-D environment. The Event-based Double Integral (EDI) [27, 28] method is proposed as a straightforward and effective way to deblur CIS frames and generate high frame rate videos. EVS and CIS measurements are connected by integrating events twice, such that the continuous intensity curve for each pixel can be determined. [34] presents a complementary filter-based framework to combine a high-pass signal from EVS and a low-pass signal from CIS to compute an all-pass signal for pixel intensity. To make the complementary filter gain adaptive to measurement noise, a pixel-wise Kalman filter is used in [39] where a unifying EVS/CIS uncertainty model is proposed.

Machine learning methods such as Time Lens [37] and its subsequent versions [10, 36, 38, 41] or others [29, 43] concatenate both CIS and EVS data using a voxel grid representation [12] as input to feed the network of two branches: an optical flow warping-based branch and a synthesis-based branch. The final result is blended with a trained attention module to avoid the disadvantages of any individual branch. CBMNet [23] is proposed to handle complex real-world motion using a novel cross-modal asymmetric bidirectional motion field estimation. Due to the high temporal resolution of EVS data, the rolling shutter problem which is commonly observed in fast-motion scenarios can be corrected using EVS data [6, 44].

EVS Latency Many studies model EVS latency [20, 25, 30, 35], but few of them have been applied to algorithm design such as frame enhancement. To the best of the authors' knowledge, this is the first work that considers an EVS latency model explicitly in generating high frame rate, deblurred, rolling-shutter artifact-free videos.

3. Problem Formulation

Event Pixel Operation and Pixel Latency Model In this work, a hybrid CIS-EVS sensor [14, 15] is used. As shown in Figure 2(a), EVS and CIS share the same photodiode design on the hybrid sensor and it is assumed that the light level is identical for nearby CIS and EVS pixels. The photocurrent i_{pd} is converted to front-end voltage V_{FE} in module `i2v`, which models dynamic characteristics as a 1st-order Low-Pass Filter (LPF) [21, 26]. The update rule for V_{FE} between discrete event indices k and $k+1$ can be written as:

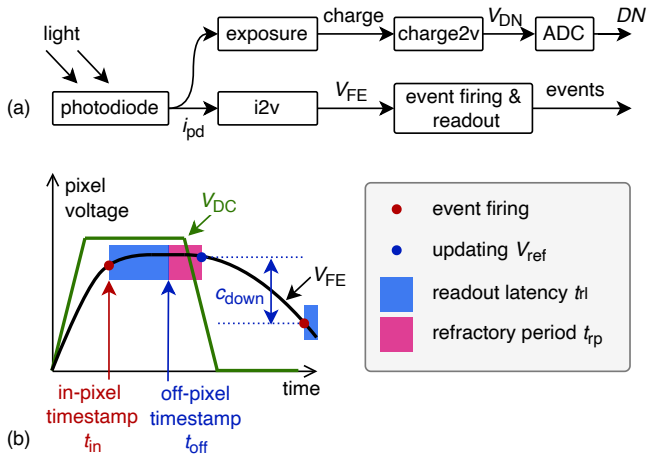

Figure 2. (a) Hybrid CIS-EVS sensor principle. The upper and lower branches indicate physical quantity conversion in CIS pixels and EVS pixels, respectively. (b) Example of V_{DC} and V_{FE} curves and triggered events that illustrate the impact of V_{FE} delay, readout latency, and refractory period.

$$V_{FE}(k+1) = V_{FE}(k) + \gamma(k)(V_{DC}(k+1) - V_{FE}(k)), \quad (1)$$

where $V_{DC}(k+1) = f_{DC}(i_{pd}(k+1))$ represents the V_{FE} under DC excitation (i.e., the V_{FE} with infinite LPF bandwidth); $f_{DC}(i_{pd})$ is the logarithmic-like current-voltage conversion function and

$$\gamma(k) = 1 - \exp\left(-\frac{\Delta t}{\tau(k)}\right) \quad (2)$$

is the time-varying coefficient of the LPF, where Δt is the time interval between two indices. The photocurrent-dependent time constant $\tau(k)$ can be modeled as [16, 26]:

$$\tau(k) = \tau_0 + \frac{\alpha}{i_{pd}(k)}, \quad (3)$$

where τ_0 and α are pixel parameters. The pixel latency in module `i2v` due to the LPF behavior is called "V_{FE} delay". As illustrated in Figure 2(b), an event is triggered at time t_{in} once the difference between V_{FE} and reference voltage V_{ref} reaches the triggering threshold voltage:

$$\begin{cases} V_{FE}(t_{in}) - V_{ref}(t_{in}) \geq c_{up} & \text{for positive events} \\ V_{FE}(t_{in}) - V_{ref}(t_{in}) \leq c_{down} & \text{for negative events} \end{cases}, \quad (4)$$

where $c_{up/down}$ denotes the contrast threshold voltage. The in-pixel timestamp t_{in} is recorded when the event fires, after which the pixel experiences "readout latency t_{rl}" and "refractory period t_{rp}". The V_{ref} is updated to follow V_{FE} after the refractory period. Finally, we have a 5-dimensional output $\{x, y, t_{in}, t_{off}, p\}$, which contains row and column indices, in-pixel and off-pixel timestamps, and polarity.

Pixel-wise Photocurrent Estimation Suppose that the illuminance on the sensor plane results in photocurrent i_{pd} for a single pixel. We aim to estimate i_{pd} when an event is triggered to avoid the impact of pixel latency in the `i2v` module. Based on Figure 2(a), we have the following CIS measurement equation:

$$DN = G \int_{t_s}^{t_e} i_{pd}(t)\mathrm{d}t, \quad (5)$$

where DN is the CIS intensity frame measured in digital numbers; G indicates the total gain of charge-voltage conversion and ADC; t_s and t_e denote the CIS exposure start and end, respectively. As for EVS measurements, suppose we have an event sequence with N events $\{x(k), y(k), t_{in}(k), t_{off}(k), p(k)\}$. By assuming $V_{FE}(1) = V_{DC}(1)$, with Eq. (1) we have the recurrence relation:

$$V_{FE}(1) = f_{DC}\left(i_{pd}(1)\right) \quad (6)$$

$$V_{FE}(k+1) = f_\gamma\left(i_{pd}(k)\right)\left(f_{DC}\left(i_{pd}(k+1)\right) - V_{FE}(k)\right), \quad (7)$$

where $k \in \{1, \cdots, N-1\}$ and $f_\gamma(\cdot)$ denotes the function that maps i_{pd} onto γ. By assuming $V_{ref}(k+1) \approx V_{FE}(k)$, the EVS measurements provide additional $N-1$ equations as:

$$c(k+1) \approx V_{FE}(k+1) - V_{FE}(k), \quad (8)$$

where $c(k) \in \{c_{up}, c_{down}\}$ are the known triggering thresholds that are configurable. These state transition equations and measurement equations (6)-(8) can be expressed as a system of $2N-1$ nonlinear equations and $2N$ unknowns - $i_{pd}(k)$ and $V_{FE}(k)$. By using a zero-order holder, the discrete states $i_{pd}(k)$ and $V_{FE}(k)$ can be approximated as continuous signals $i_{pd}(t)$ and $V_{FE}(t)$. This makes it possible to utilize CIS measurements in Eq. (5) as additional constraints. In conclusion, the pixel latency compensation can be expressed as pixel-wise optimization problem, solving the nonlinear system with $2N$ unknown states and $2N+M-1$ equations, where M and N denote the number of CIS and EVS measurements, respectively.

Modeling Readout Latency and Refractory Period The EVS measurement model equation (8) is an approximation since the state variation during the readout latency and the refractory period is not considered. Using both in-pixel timestamps and off-pixel timestamps, it is possible to compensate for some of the "missed" V_{FE} change by assuming that the slope of V_{FE} is locally constant for a short period. As illustrated in Figure 2(b), it is noted that:

$$m(k+1) = \frac{c(k+1)}{t_{in}(k+1) - t_{off}(k) - t_{rp}}, \quad (9)$$

where m describes the slope approximation of the V_{FE} curve. For the readout latency, the missed V_{FE} is represented as:

$$\Delta V_{FE}^{rl}(k) = m(k+1)\left[t_{off}(k) - t_{in}(k)\right]. \quad (10)$$

For the refractory period, we have $\Delta V_{FE}^{rp}(k) = m(k+1)t_{rp}$ since t_{rp} is a fixed parameter. Thus the reference voltage after compensation can be adjusted to:

$$V_{ref}(k+1) = V_{FE}(k) + \Delta V_{FE}^{rl}(k) + \Delta V_{FE}^{rp}(k). \quad (11)$$

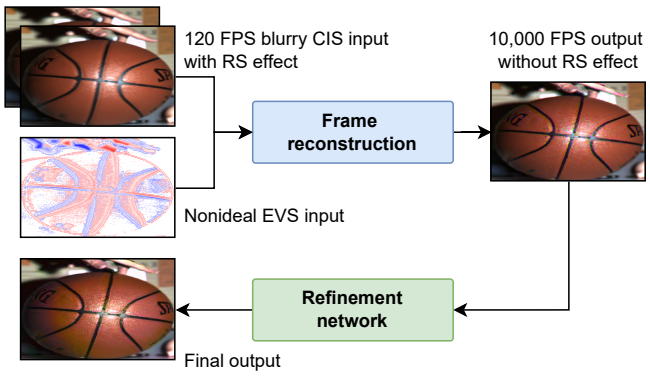

Figure 3. Block diagram of the proposed method. Note that after the frame reconstruction the blurry texture due to motion is enhanced and the RS effect is removed. The noise and artifacts are further reduced by the refinement network.

By substituting the compensation equation (11) into (4), the impact of readout latency and refractory period can be compensated:

$$c(k+1) \approx V_{\mathrm{FE}}(k+1) - V_{\mathrm{FE}}(k) - \Delta V_{\mathrm{FE}}^{\mathrm{sl}}(k) - \Delta V_{\mathrm{FE}}^{\mathrm{rp}}(k), \quad (12)$$

where $\Delta V_{\mathrm{FE}}^{\mathrm{rl}}(k)$ and $\Delta V_{\mathrm{FE}}^{\mathrm{rp}}(k)$ are voltage compensation for readout latency and refractory period, respectively. Note that this compensation is enabled by the availability of in-pixel and off-pixel timestamps.

4. Method

The proposed frame enhancement framework is presented in Figure 3. The frame reconstruction module jointly deblurs images, removes the RS effect, and interpolates high frame rate images. The refinement network focuses on removing noise and improving image quality.

4.1. Solving the Inverse Model via Optimization

Graph Modeling for a Single Pixel Due to the presence of pixel latency, the system is governed by nonlinear equations, rendering the models and solutions designed for linear systems inapplicable. Graph optimization [13] provides a straightforward way to model nonlinear systems, where the states and measurements are represented as nodes and edges in a graph as shown in Figure 4(a). Let $l \in \{1, \cdots, M\}$ and $k \in \{1, \cdots, N-1\}$ be the indices of CIS and EVS, respectively. By defining states $\boldsymbol{x} = [i_{\mathrm{pd}}(k)]^\top$ and measurements $\boldsymbol{z}_{\mathrm{CIS}} = [DN(l)]^\top$, $\boldsymbol{z}_{\mathrm{EVS}} = [c(k+1)]^\top$, the pixel-wise graph optimization framework aims to determine the state vector $\boldsymbol{x}$:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left(\boldsymbol{e}_{\mathrm{CIS}}^\top \boldsymbol{\Omega}_{\mathrm{CIS}} \boldsymbol{e}_{\mathrm{CIS}} + \boldsymbol{e}_{\mathrm{EVS}}^\top \boldsymbol{\Omega}_{\mathrm{EVS}} \boldsymbol{e}_{\mathrm{EVS}} \right) \quad (13)$$

where $\boldsymbol{e}_{\mathrm{CIS}} = \boldsymbol{z}_{\mathrm{CIS}} - \tilde{\boldsymbol{z}}_{\mathrm{CIS}}$, $\boldsymbol{e}_{\mathrm{EVS}} = \boldsymbol{z}_{\mathrm{EVS}} - \tilde{\boldsymbol{z}}_{\mathrm{EVS}}$ denote measurement error vectors; $\tilde{\boldsymbol{z}}_{\mathrm{CIS}}$ and $\tilde{\boldsymbol{z}}_{\mathrm{EVS}}$ are predicted measurements which can be calculated from estimated states according to CIS measurement equation (5) and EVS measurement equation (8); $\boldsymbol{z}_{\mathrm{EVS}}$ is adjusted using register settings and calibrated using the "S-curve method "[7]. The statistical distribution of the threshold is not explicitly modeled in the current method. The diagonal weighing matrices $\boldsymbol{\Omega}_{\mathrm{CIS}}$ and $\boldsymbol{\Omega}_{\mathrm{EVS}}$ reflect the reliability of CIS and EVS, respectively. Using the cost function equation (13), the V_{FE} delay is modeled explicitly through the EVS measurement equations. Solving the optimization problem leads to state estimation such that the V_{FE} delay is compensated. Many approaches have been proposed to solve nonlinear optimization problems iteratively [19]. We parallelize pixel-wise tasks as there is no data interaction or computational dependency between pixels.

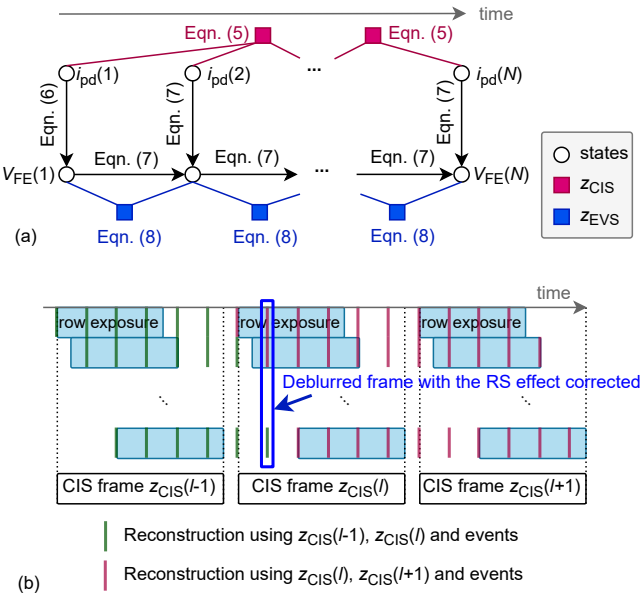

Figure 4. (a) Graph modeling of the pixel-wise photocurrent estimation problem, where nodes and edges denote states and measurements. The arrows indicate the order of state calculation. "$\boldsymbol{z}_{\mathrm{CIS}}$" and "$\boldsymbol{z}_{\mathrm{EVS}}$" indicate CIS and EVS measurements, respectively. Eqn. (8) is replaced with Eqn. (12) when voltage compensation for readout latency and refractory period is enabled. (b) Row-based reconstruction scheme showing three CIS images as input. After row-wise computation, interpolated and deblurred frames without RS effect can be generated as shown in the blue box. The proposed scheme can be applied to video enhancement with any number of CIS images.

Rolling Shutter (RS) Effect Correction As CIS parameters such as row exposure time and RS scanning speed are known, we formulate the frame estimation problem as pixel-wise computation using adjusted CIS timing parameters for each row. After solving the photocurrent at each event time-step through optimization, a zero-order hold is used to generate a time-continuous photocurrent $i_{\mathrm{pd}}(t)$ for all pixels. The CIS measurement equation (5) is used to generate a rolling shutter artifact-free target frame rate (such

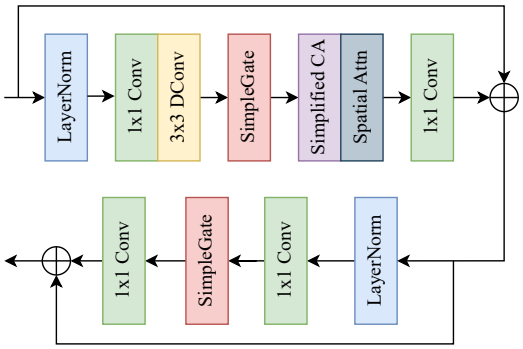

Figure 5. Modified NAF block in the refinement network.

as 10 000 FPS).

Video Glitch Effect Deblurring and video frame interpolation performance can experience glitches in case these are computed on a frame-by-frame basis. To mitigate this issue, we solve the coupled deblurring, rolling-shutter correction and interpolation problem jointly for subsequent frames utilizing the row-specific rolling shutter timing characteristics as shown in Figure 4(b), where we reconstruct each row according to its specific exposure time by fusing two consecutive CIS frames and their corresponding events.

4.2. Refinement Network

A refinement network is developed as a post-processing module to further mitigate the noise and artifacts in the reconstructed frames. NAFNet [4] is designed for image restoration. It is structured as a U-Net, arranging NAF blocks to allow the network to progressively downsample the input for semantic feature extraction and then progressively upsample it to restore the output. NAFNet provides additional skip connections that preserve information and allow feature reusability from earlier layers. NAF blocks are structured as original ResNet blocks, with additional layers such as layer normalization and channel attention. We adopt the NAFNet basic concept, and further modify making two major additions: 1) spatial attention [40] as shown in Figure 5 and 2) perceptual loss [32]. The reason for choosing spatial attention is that most of the noise is located in motion areas instead of being ubiquitous across the image. Spatial attention allows the network to focus on these noisy regions of the image, which contributes to better denoising results. Moreover, the rationale for choosing the perceptual loss is that even though the proposed network does well in removing the pixel-level saturation noise, it also introduces blurriness or texture degradation. Perceptual loss helps to recover the image quality in a way that better aligns with the human vision system. The losses applied in this paper are summarized as:

$$\mathcal{L}_{\text{s}} = w_l \mathcal{L}_1 + w_{\text{VGG}} \mathcal{L}_{\text{VGG}} + w_{\text{Gram}} \mathcal{L}_{\text{Gram}}, \qquad (14)$$

where the $\mathcal{L}_1$ loss is the pixel-wise difference between the reconstructed RGB image and ground truth image; the $\mathcal{L}_{\text{VGG}}$ loss is calculated as $\mathcal{L}_{\text{VGG}} = \frac{1}{L}\sum_{l=1}^{L} \alpha_l \left\| \boldsymbol{\Psi}_{\boldsymbol{l}}(\boldsymbol{I}_{\text{Recon}}) - \boldsymbol{\Psi}_{\boldsymbol{l}}(\boldsymbol{I}_{\text{GT}}) \right\|_1$; $\boldsymbol{\Psi}_{\boldsymbol{l}}(\boldsymbol{I})$ is the feature tensor from the l_{th} layer of an ImageNet pre-trained VGG-19 network generated by passing the image $\boldsymbol{I}$ through the network. In this case, L specifies the total number of layers in the VGG-19 network, while α_l is the hyperparameter assigned to each layer. Finally, the Gram loss [11] is computed as the L2 difference between the autocorrelation of the VGG-19 features of the reconstructed image and ground truth image with the formulation: $\mathcal{L}_{\text{Gram}} = \frac{1}{L}\sum_{l=1}^{L} \alpha_l \left\| \boldsymbol{M}_{\boldsymbol{l}}(\boldsymbol{I}_{\text{Recon}}) - \boldsymbol{M}_{\boldsymbol{l}}(\boldsymbol{I}_{\text{GT}}) \right\|_2$, where given an image I, the Gram matrix is computed for every layer l as $\boldsymbol{M}_l = (\boldsymbol{\Psi}_l(\boldsymbol{I}))^{\top}(\boldsymbol{\Psi}_l(\boldsymbol{I}))$.

5. Experiments

We compare our proposed method with state-of-the-art EVS+CIS video reconstruction methods: EDI [27], AKF [39], Time Lens [37], CBMNet [23], EvUnroll [8] and REFID [36]. Among them, EvUnroll proposes an RS correction based reconstruction and REFID combines image deburring with EVS based frame interpolation. We also show the results of the CIS-only interpolation method FILM [32] as reference. Before optimization, i_{pd} are initialized according to the first CIS image DN and the inverse photocurrent-DN relation. The elements in the weighing matrices for CIS, EVS terms are set to 1. The performance of the interpolation methods are evaluated both quantitatively and qualitatively, based on the metrics of Peak Signal-to-Noise Ratio (PSNR), LPIPS [42], and pixel-wise reconstruction error on DN. Furthermore, to study the performance degradation concerning ambient illuminance and motion speed, we measured a test set utilizing a rotating disk in a controlled environment. We analyze the Blurred Edge Width (BEW) [5] of the reconstructed frames – a larger BEW indicates a blurrier image and, therefore, a worse deblurring result.

5.1. Datasets

We evaluate the proposed method using the **simulation data** based on the CIS-EVS hybrid sensor simulator [26] which considers V_{FE} delay, readout latency, and refractory period. Our **measurement data** was collected from the hybrid CIS-EVS sensor [14, 15]. Here, only off-pixel timestamp data was captured. Two measured datasets were created for evaluation: The natural scene dataset focuses on indoor scenarios where EVS is significantly influenced by pixel latency. The other dataset captured a rotating disk with a Siemens star pattern under different ambient illumination levels and rotating speeds. In addition to the above datasets specifically captured for this work's evaluation, we also test our method using the publicly available HS-ERGB

Table 1. Quantitative comparison of state-of-the-art video reconstruction methods on the proposed simulation dataset in terms of PSNR (in dB, higher is better) and LPIPS (a dimensionless quantity, lower is better). Each row shows results for a particular scene. Results from different methods are listed column-wise. The first and second places are highlighted with **bold underline** and **bold**, respectively.

	Ours w/o refinement		Ours		EDI		AKF		Time Lens		FILM		CBMNet		EvUnroll		REFID	
Scene	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
basketball	24.01	**0.217**	**27.23**	**0.155**	22.49	0.220	15.63	0.280	23.65	0.240	23.86	0.220	**24.51**	0.252	23.13	0.218	22.38	0.223
checkerboard	**24.23**	**0.203**	**28.24**	**0.187**	20.48	0.218	19.56	0.261	22.60	0.239	23.02	0.216	22.76	0.212	23.84	0.204	21.23	0.235
slingshot egg	**25.23**	**0.100**	**32.85**	**0.106**	24.95	0.164	15.59	0.110	25.19	0.244	24.04	0.203	22.76	0.234	24.61	0.210	23.92	0.201
running man	24.60	0.164	24.72	0.189	22.29	0.135	20.71	0.189	25.01	0.117	**25.06**	**0.089**	**25.16**	**0.069**	23.04	0.117	23.52	0.163
fan	**22.40**	0.189	**24.50**	**0.163**	18.17	0.255	10.78	**0.184**	20.64	0.233	18.93	**0.184**	21.28	0.204	19.64	0.190	18.30	0.198
Average	**24.09**	**0.175**	**27.51**	**0.16**	21.68	0.198	16.45	0.205	23.42	0.215	22.98	0.182	23.29	0.194	22.85	0.188	21.87	0.204

Table 2. Quantitative results of ablation study. The first and second places are highlighted with **bold underline** and **bold**, respectively.

	without compensation		with V_{FE} delay compensation		with RL+RP compensation		with V_{FE}+RL+RP compensation		with V_{FE}+RL+RP compensation and refinement	
Scene	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
basketball	22.95	0.316	23.47	0.306	23.50	0.307	**23.56**	**0.302**	**23.64**	**0.295**
slingshot egg	23.57	**0.099**	**24.00**	**0.099**	23.60	**0.099**	23.71	0.102	**29.04**	0.106

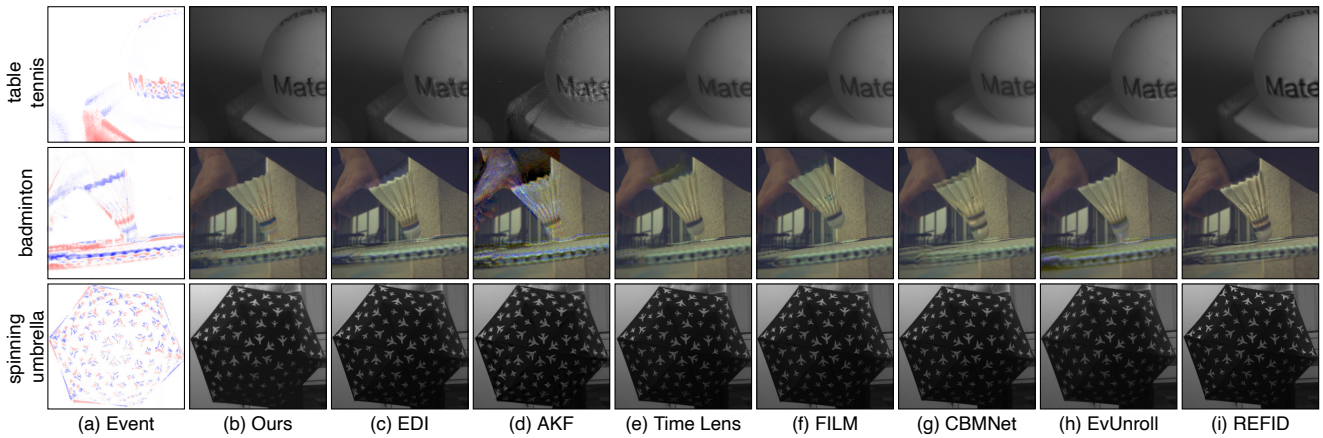

Figure 6. Qualitative comparison of state-of-the-art video reconstruction methods on the proposed measured datasets in terms of image quality. The images have been magnified for optimal viewing quality.

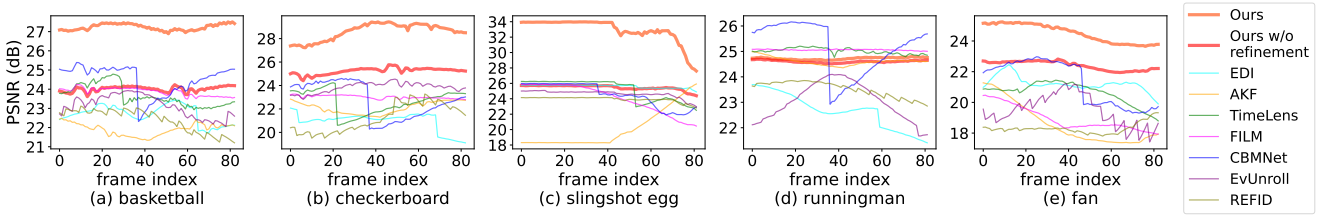

Figure 7. Frame-wise, quantitative comparison of state-of-the-art video reconstruction methods in terms of PSNR for varying scenarios of the simulation dataset. Note that some algorithms exhibit pronounced discontinuities in the observed PSNR curves indicating the presence of video glitches.

dataset based on a EVS camera that is collocated next to a CIS camera. Since the paper focuses on mitigating artifacts due to pixel latency, two indoor scenarios ("spinning plate" and "spinning umbrella") were selected. More details of the dataset used are provided in the supplementary material.

5.2. Results

Image Quality Table 1 shows quantitative results for the reconstructed video. Results in PSNR and LPIPS show that the proposed method outperforms other state-of-the-art methods in most of the scenes, especially with fast motion ("slingshot egg") or complicated rotation motion ("basket-

ball", "fan"). As expected the refinement network which aims to alleviate artifacts and improve image quality reports a higher score. Snippets of the reconstructed videos are presented in Figure 1 and Figure 6 for simulated and measured data, respectively. Comprehensive input images are provided in the supplementary material. Thanks to the explicit EVS latency modeling and compensation, our results avoid obvious shadowing/ghosting artifacts – especially if there is fast motion under challenging lighting conditions. It is also noted that the proposed method successfully corrects the RS effect, as shown in the "basketball" scene. Learning-based methods face difficulties, as latency and the RS effect have not been considered in model design and training. The CIS-only method FILM provides natural image quality with low noise but fails when the objects are too fast to be captured by CIS, as shown in the "slingshot egg" scene. For the "basketball" scene, our method shows an advantage in details and sharpness compared to FILM.

In specific scenarios, such as the "spinning umbrella" scene within the HS-ERGB dataset and the "running man" scene in the simulation dataset, our method did not yield improved image quality. This observation can be attributed to several factors. Firstly, the HS-ERGB dataset employs a different EVS with different pixel and readout latency as well as refractory period. Secondly, this dataset is based on a dual camera instead of a hybrid sensor so the spatio-temporal alignment may differ. Thirdly, the scenes in question involve relatively slower object movement speeds resulting in reduced ghosting. In essence, the dynamic characteristics of the EVS, the input signal frequency range, and the reconstruction algorithm influence the quality of the reconstructed images.

Video Glitch Effect To evaluate the glitch issue, we compute the PSNR for each reconstructed frame as shown in Figure 7. Overall, our method gives temporally more stable PSNR curves, which stem from the proposed row-based reconstruction scheme. It is also noted that the refinement network contributes significantly to higher PSNR.

Pixel-wise Error Analysis We select a pixel within the region of interest (indicated as the yellow box in Figure 1(b)) and verify the pixel-wise reconstruction accuracy. As shown in Figure 8(a), with the help of pixel latency compensation the reconstructed V_{DC} after optimization shows a sharper peak (with 250 µs duration) compared to V_{FE} (with over 1 ms duration). The high-frequency fluctuations on ground truth V_{DC} curve cannot be reconstructed due to the V_{FE} bandwidth limit. An overshoot is observed after the peak resulting in a dark tail along the pellet's trajectory, as shown in Figure 1(b). This may stem from 1) the V_{FE} bandwidth, 2) the approximation error in equation (12), 3) EVS measurement noise, and 4) choice of weighing matrices for CIS and EVS. Figure 8(b), indicates that the proposed method gives an excellent reconstruction of the pellet's peak DN despite the overshoot.

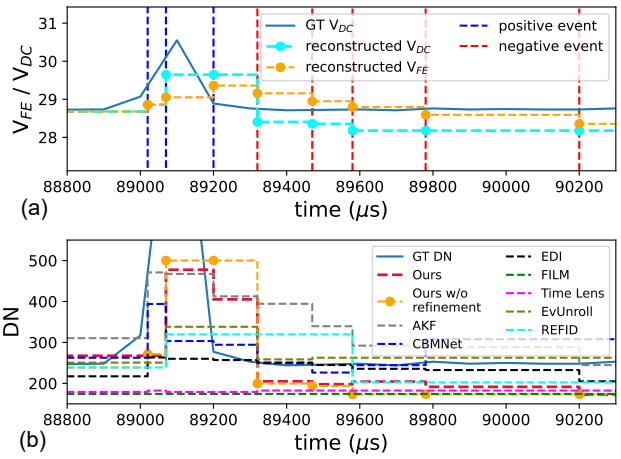

Figure 8. Pixel-wise reconstruction results for the "slingshot egg" scene. The selected pixel is located at the path of the pellet, in the yellow box of Figure 1. The refinement network is not used here. (a) Comparison of the ground truth V_{DC}, reconstructed V_{FE} and V_{DC}, where the ground truth V_{DC} is computed from the ground truth DN of the pixel at 10 000 FPS. (b) Comparison of reconstructed pixel-wise DN of state-of-the-art video reconstruction methods.

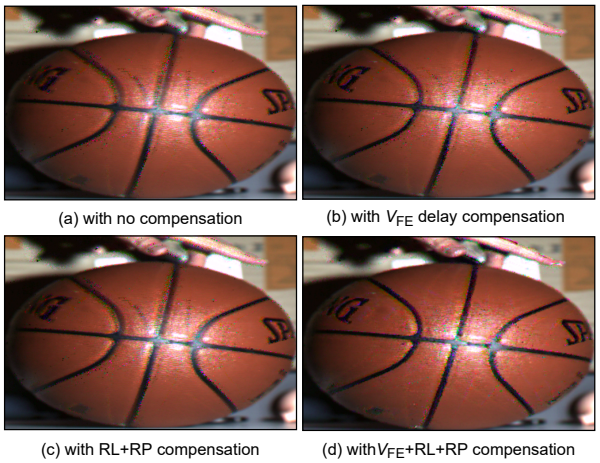

Figure 9. Qualitative results of ablation study on the "basketball" scene. Refer to Figure 1(a) for "with V_{FE}+RL+RP compensation and refinement" result.

Ablation Study We ablate the refinement network, V_{FE} delay compensation, readout latency (RL) and refractory period (RP) compensation of our method. The numerical evaluation are subject to selected ROIs on motion region. To ablate V_{FE} delay compensation, we set the time constant $\tau = 0$ such that $V_{\mathrm{FE}} = V_{\mathrm{DC}}$ in Eq. (7). To ablate readout latency and refractory period compensation, we use Eq. (8) instead of Eq. (12) in the EVS measurement model. Results are summarized in Table 2 and Figure 9. The con-

figuration using all modules gives significantly better visual and slightly better numerical results in PSNR compared to having latency compensation and refinement disabled. Disabling V_{FE} or RL+RP compensation significantly worsens the image quality, as severe shadowing artifacts are observed. As for LPIPS in "slingshot egg" scene, the difference is not significant due to the small area of pellet ROI. It is also noted that adding RL+RP compensation leads to improvement in the "basketball" but degrades the performance in "slingshot egg" when comparing with V_{FE} delay compensation only. This is because RL+RP compensation is based on the assumption that the slope of VFE is locally constant for a short period. In cases such as "slingshot egg", the assumption may not hold due to close positive and negative events. An ablation study against the use of NAFNet-only is included in supplementary material.

Performance Regarding Illuminance and Speed The proposed method is further investigated under a controlled environment using varying illuminance and object motion speed. BEW [5] is used to evaluate the image quality by measuring edge sharpness. Figure 10(a)-(h) presents the comparison of the state-of-the-art methods on a rotating disk under 1000 lx and 292 rpm. It is observed that the proposed method reconstructs sharper edges compared to the others. Other EVS-based methods struggle resolving blurry edges in presence of fast motion and also the CIS-based FILM method fails to generate a sharp reconstruction. Figure 10(i) shows that the proposed method reports the lowest BEW score (0.7) and has the sharpest edge profile.

The BEW scores under different speed and illumination conditions are summarized in Figure 11. Figure 11(a) shows that the BEW increases with faster rotation speeds. In Figure 11(b), the lower ambient light condition leads to higher V_{FE} latency degrading image quality. Our method achieves the lowest BEW throughout all illuminance levels.

6. Conclusion

An EVS-assisted framework to solve motion blur, rolling-shutter artifacts and video-frame interpolation in a joint manner was presented. Existing methods use either a simple double integral model or a learning-based model, neither of which explicitly address EVS sensor imperfections such as EVS pixel latency, readout latency or refractory period. This work overcomes these limitations by explicitly modeling these characteristics. The compensation of readout latency and refractory period is enabled by having access to in-pixel and off-pixel time-stamps. The nonlinear inverse model was solved through joint graph optimization of subsequent frames utilizing row-by-row exposure information to overcome rolling-shutter artifacts and possible glitches between subsequent frames. A refinement network was proposed to further improve the image quality by using specifically designed spatial attention blocks. We showed that

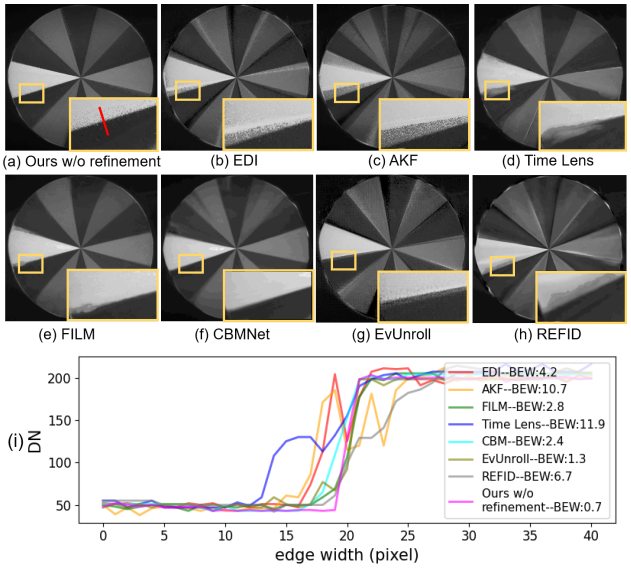

Figure 10. (a)-(h) Comparison of state-of-the-art video reconstruction methods on the proposed rotating disk datasets with rotating speed 292 rpm and environment illuminance 1000 lx. The yellow box shows magnified edges. The visual "misalignment" is due to EVS latency. (i) The DN-edge width relation at a fixed segment of 40 pixels (red line in (a)) on reconstructed rotating disk edge. The refinement network is not used.

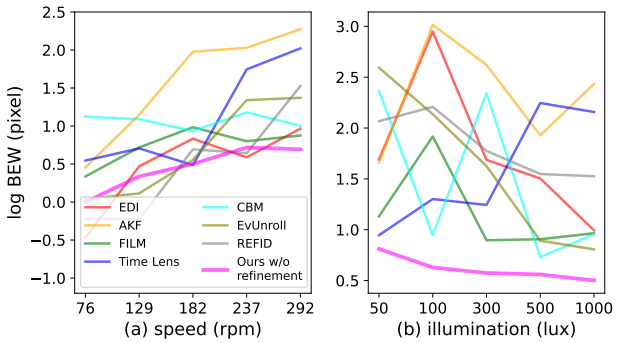

Figure 11. Quantitative comparison of state-of-the-art video reconstruction methods on the proposed "rotating disk" scene in terms of BEW. The lower the BEW values, the sharper the edge. The refinement network is not used. (a) BEW under various rotation speeds with a fixed illuminance of 1000 lx. (b) BEW under various illuminance levels with a fixed speed of 292 rpm.

the proposed method outperforms state-of-the-art methods in reconstructed image quality with an up to 4 dB improvement in PSNR and 12 % improvement in LPIPS score.

Limitation Since the pixel latency highly depends on EVS design parameters, object motion, and ambient illuminance, the image quality improvement after latency compensation may be subtle for slow objects or in bright scenes such as "running man" and "spinning umbrella".

References

[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3703–3712, 2019. 2

[2] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 884–892, 2016. 2

[3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1

[4] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1239–1248, 2022. 5

[5] Hai Dinh, Qinyi Wang, Fangwen Tu, Brett Frymire, and Bo Mu. Evaluation of motion blur image quality in video frame interpolation. *Electronic Imaging*, 35:262–1, 2023. 5, 8

[6] Julius Erbach, Stepan Tulyakov, Patricia Vitoria, Alfredo Bochicchio, and Yuanyou Li. Evshutter: Transforming events for unconstrained rolling shutter correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13904–13913, 2023. 2

[7] C.Posch et al. Sensitivity and uniformity of a 0.18µm CMOS temporal contrast pixel array. In *ISCAS*, 2011. 4

[8] Xinyu Zhou et al. Evunroll: Neuromorphic events based rolling shutter image correction. In *CVPR*, 2022. 2, 5

[9] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1

[10] Yue Gao, Siqi Li, Yipeng Li, Yandong Guo, and Qionghai Dai. Superfast: 200× video frame interpolation via event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 5

[12] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 2

[13] Giorgio Grisetti, Rainer Kümmerle, Hauke Strasdat, and Kurt Konolige. g2o: A general framework for (hyper) graph optimization. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 9–13, 2011. 4

[14] Menghan Guo, Shoushun Chen, Zhe Gao, Wenlei Yang, Peter Bartkovjak, Qing Qin, Xiaoqin Hu, Dahai Zhou, Qiping Huang, Masayuki Uchiyama, Yoshiharu Kudo, Shimpei Fukuoka, Chengcheng Xu, Hiroaki Ebihara, Xueqing Wang, Peiwen Jiang, Bo Jiang, Bo Mu, Huan Chen, Jason Yang, T. J. Dai, and Andreas Suess. A three-wafer-stacked hybrid 15-mpixel cis + 1-mpixel evs with 4.6-gevent/s readout, in-pixel tdc, and on-chip isp and esp function. *IEEE Journal of Solid-State Circuits*, pages 1–10, 2023. 1, 2, 5

[15] Menghan Guo, Shoushun Chen, Zhe Gao, Wenlei Yang, Peter Bartkovjak, Qing Qin, Xiaoqin Hu, Dahei Zhou, Masayuki Uchiyama, Yoshiharu Kudo, Shimpei Fukuoka, Chengcheng Xu, Hiroaki Ebihara, Andy Wang, Peiwen Jiang, Bo Jiang, Bo Mu, Huan Chen, Jason Yang, TJ Dai, and Andreas Suess. A 3-wafer-stacked hybrid 15MPixel CIS + 1MPixel EVS with 4.6GEvent/s readout, in-pixel TDC and on-chip ISP and ESP function. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, 2023. 2, 5

[16] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1312–1321, 2021. 3

[17] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642. Springer, 2022. 2

[18] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9000–9008, 2018. 2

[19] Nocedal Jorge and J Wright Stephen. *Numerical optimization*. Spinger, 2006. 4

[20] Damien Joubert, Mathieu Hébert, Hubert Konik, and Christophe Lavergne. Characterization setup for event-based imagers applied to modulated light signal detection. *Applied optics*, 58(6):1305–1317, 2019. 2

[21] Damien Joubert, Alexandre Marcireau, Nic Ralph, Andrew Jolley, André van Schaik, and Gregory Cohen. Event camera simulator improvements via characterized parameters. *Frontiers in Neuroscience*, page 910, 2021. 1, 2

[22] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 349–364. Springer, 2016. 2

[23] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18032–18042, 2023. 1, 2, 5

[24] Kazutoshi Kodama, Yusuke Sato, Yuhi Yorikado, Raphael Berner, Kyoji Mizoguchi, Takahiro Miyazaki, Masahiro Tsukamoto, Yoshihisa Matoba, Hirotaka Shinozaki, Atsumi Niwa, et al. 1.22 μm 35.6 mpixel rgb hybrid event-based vision sensor with 4.88 μm-pitch event pixels and up to 10k event frame rate by adaptive control on event sparsity.

In *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 92–94. IEEE, 2023. 1

[25] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2):566–576, 2008. 1, 2

[26] Xiaozheng Mou, Kaijun Feng, Alex Yi, Steve Wang, Huan Chen, Xiaoqin Hu, Menghan Guo, Shoushun Chen, and Andreas Suess. Accurate event simulation using high-speed videos. *Electronic Imaging*, 34(7):242–1, 2022. 2, 3, 5

[27] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2019. 1, 2, 5

[28] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2519–2533, 2020. 2

[29] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 2

[30] Christoph Posch and Daniel Matolin. Sensitivity and uniformity of a 0.18 μm cmos temporal contrast pixel array. In *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, pages 1572–1575. IEEE, 2011. 2

[31] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 2

[32] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision*, pages 250–266. Springer, 2022. 1, 2, 5

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[34] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, pages 308–324. Springer, 2018. 1, 2

[35] Andreas Suess, Menghan Guo, Rui Jiang, Xiaozheng Mou, Qiping Huang, Wenlei Yang, and Shoushun Chen. Physical modeling and parameter extraction for event-based vision sensors. In *Proc. IISS Int. Image Sensor Workshop (IISW), Edinburgh, UK*, page R5, 2023. 2

[36] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 1, 2, 5

[37] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021. 1, 2, 5

[38] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 2

[39] Ziwei Wang, Yonhon Ng, Cedric Scheerlinck, and Robert Mahony. An asynchronous kalman filter for hybrid event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 448–457, 2021. 1, 2, 5

[40] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 5

[41] Zeyu Xiao, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Eva^2: Event-assisted video frame interpolation via cross-modal alignment and aggregation. *IEEE Transactions on Computational Imaging*, 8:1145–1158, 2022. 2

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[43] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17774, 2022. 2

[44] Xinyu Zhou, Peiqi Duan, Yi Ma, and Boxin Shi. Evunroll: Neuromorphic events based rolling shutter image correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17775–17784, 2022. 1, 2